近代报纸资源细粒度语义描述模型设计及应用*
——以《盛京时报》为例

孙绍丹[1] 邓君[1] 常严予[1] 张子姝[1] 沈涌[2]
[1] 吉林大学商学与管理学院 长春 130012 [2] 吉林大学公共卫生学院 长春 130022

摘 要：[目的/意义]设计科学、规范的近代报纸资源细粒度语义描述模型，深入揭示近代报纸资源特征及关联关系，为近代报纸资源的有效管理、组织及知识发现、知识服务提供参考。[方法/过程]通过分析近代报纸资源逻辑结构、物理布局、内容信息等，从领域本体和元数据描述两方面着手，复用 CIDOC-CRM 本体概念模型和 EAD、DC、《古籍元数据规范》，以《盛京时报》为例设计近代报纸资源细粒度语义描述模型，并采用 Oxygen XML 工具将语义描述模型用 RDF/XML 语言进行描述，实现元素互操作和模型应用。[结果/结论]为近代报纸资源组织提供一个可操作的细粒度语义描述模型，为近代报纸资源库构建、报纸规范化管理及应用系统开发等提供基础保障，促进近代报纸资源的开发、利用与共享。

关键词：近代报纸资源 语义描述模型 《盛京时报》 RDF/XML
分类号：G254
DOI：10.13266/j.issn.0252-3116.2022.07.004

1 引言

近代报纸文献全方位记载了近代社会的巨大变革，承载着时代印记，是研究社会史、政治史、经济史、文化史、新闻史的重要信息来源。近代报纸起源于西方资本主义萌芽阶段，如德国人古登堡在 15 世纪下半叶发明了金属活版印刷术，随之报纸雏形“新闻书”出现；17 世纪资本主义在欧洲盛行，各国代表性报纸日益涌现。较于西方，中国近代报纸诞生较晚，第一批由外国传教士在华创办，如《察世俗每月统纪传》《东西洋考每月统纪传》《申报》及《盛京时报》等。这些报纸由外国人创办，带有主观的政治倾向色彩，以传播西方文化为主。近代国人最早自办的日报是艾小梅于 1873 年在汉口创办的《昭文新报》[1]，继之戊戌变法、辛亥革命和新文化运动发生，改良派、革命派及新型资产阶级逐渐登上历史舞台，《时务报》《知新报》《国闻报》等报纸也陆续涌现，成为舆论宣传阵地。五四运动作为中国现代史开端，将马克思主义成功引入国内，革命报纸《向导》《新青年》《共产党》《中国青年》等创刊发行，政治功能凸显。此后，土地革命、抗日战争和解放战争爆发，红色报纸大批刊行，如《红星》《抗战日报》《解放日报》《晋察冀日报》。由此，中国近代报刊主要以救亡图存为目标，围绕“启蒙”和“革命”展开[2]，且该时期（1840－1949 年）报纸栏目多样，内容包罗万象，上至国内外重大时政新闻动态，下至市井民众生活百态，报道内容精细入微，是中国近代历史画卷的生动“缩影”，在社会变革和历史动荡中扮演着重要角色，具有珍贵的史料价值、学术价值和历史史实的订正价值[3]。

近年来，新一轮科技革命蓬勃兴起，数字技术更新迭代，文献数字化成为历史发展的必然趋势。近代报纸资源在技术赋能下迎来了新的发展机遇。然而，由于近代报纸距今久远，纸张老旧、酸化脆弱、破损严重，面临严峻的保存危机。为了抢救和保护近代报纸资源，弘扬优秀传统文化，一些学术组织和商业机构等借助现代信息技术实现近代报纸的数字化长期保存及利用。如以国家图书馆为引领的文化单位启动了“革命文献与民国时期文献保护计划”项目，囊括了大批量近

* 本文系国家社会科学基金项目“数字人文视角下历史档案资源知识聚合与知识发现研究”（项目编号：19BTQ102）研究成果之一。
作者简介：孙绍丹，博士研究生；邓君，教授，中国人民大学档案事业发展研究中心研究员，博士，博士生导师，通信作者，E-mail：dengjun9722@163.com；常严予，硕士研究生；张子姝，博士研究生；沈涌，讲师。
收稿日期：2021－09－26 修回日期：2021－11－28 本文起止页码：35－46 本文责任编辑：王传清

代报纸文献,该项目受到中央政府的高度重视。2016 年该项目被列入《中华人民共和国国民经济和社会发展第十三个五年规划纲要》,2017 年又被写入《国家"十三五"时期文化发展改革规划纲要》和《文化部"十三五"时期文化发展改革规划》。此外,其他国家图书馆也以工具书编制、专题报纸数据库构建、影印出版、缩微复制、数字加工等方式进行历史报纸数字化活动,为用户提供报纸资源浏览和检索功能。如美国国会图书馆数字报纸计划、欧洲报纸数字化项目、芬兰图书馆数字报纸项目、澳大利亚国家图书馆数字报纸项目等。这些项目推动了报纸资源数字化建设进程,为报纸资源高效开发和利用奠定了基础。

然而,笔者通过调研国内一些近代报纸资源库及索引库,发现近代报纸资源库缺少统一规范的报纸语义描述模型作为支撑,往往是文本图片等资源的简单堆砌和陈列,多个资源库检索方式单一化,缺乏对报纸内容的深度揭示及多维语义关联关系的挖掘与组织,停留在物理载体层面的简要描述,严重限制了用户获取细粒度信息的多种可能性,也无法快速锁定目标需求资源,影响用户服务质量。因此,有必要设计一个全面、规范、可互操作的近代报纸资源细粒度语义描述模型,抽取结构化知识以满足用户的复杂信息获取和检索需求,提高知识服务效率。在此背景下,本文依据近代报纸逻辑结构、物理布局及内容特征,从领域本体和元数据描述入手,设计近代报纸资源细粒度语义描述模型,以期为近代报纸资源库构建、报纸规范化管理及应用系统开发等提供基础保障,促进近代报纸资源的开发、利用与共享。

2 相关研究

通过对国内外近代报纸资源相关研究文献的梳理,发现学者们主要聚焦于报纸数字化项目建设、报纸抢救及长期保存、报纸数字化过程中数据质检、数据噪音等问题以及报纸资源知识组织研究等 4 个层面。

2.1 报纸数字化项目建设方面

P. Tonijala 等全面细致地介绍了美国国家数字报纸计划项目,并提出将报纸资源内容嵌入到教育教学过程中[4];R. Atanassova 等对欧洲图书馆数字报纸项目网站建设历程及功能模块进行分析[5-6],以满足数字人文研究者相关知识需求。国内学者主要以国家图书馆、北京大学图书馆、首都图书馆、上海图书馆或区域省市图书馆所收藏的近代报纸等为例,从数字化报纸品种、报纸资源数量、报纸资源类型、时间范围、内容选择、检索和阅读功能、报纸资源服务方式等层面解析民国报纸资源建设现状[7],并提出相应的解决对策和建议[8]。

2.2 报纸抢救及长期保存方面

A. Krahmer 以北德克萨斯大学和斯坦福大学的合作项目 The Texas Digital Newspaper Program (TDNP) 为例,阐述报纸数字化保存策略[9];M. Georgieva 以内华达州数字报纸项目为例,从项目管理视角讨论报纸项目管理技术和工具、如何进行报纸抢救及长期保存策略[10]。国内学者则以地方近代报纸数字化建设为例[11],探讨其数字化技术和工具、分析数字化报纸的必要性和优势,并提出抢救近代报纸的相关建议[12]。

2.3 报纸数据质检方面

J. Jarlbrink 等分析了瑞典国家图书馆在历史报纸数字化过程中的数字噪音问题[13],如光学字符识别(Optical Character Recognition, OCR)识别质量参差不齐、载体形态转换价值丢失、数字外包质量控制风险等。数字噪音是报纸数字化过程中的焦点问题,数字化质量直接影响报纸资源的开发利用;国内学者探讨了民国报纸数字化实践中的质检问题,如报纸版式识别和 OCR 文字识别、报纸记录标识号、报名、出版日期、版次、栏目等问题[14]。

2.4 报纸知识组织方面

学者们主要围绕报纸资源元数据描述规范展开论述。在报纸数字化实践项目中,一般采用书目元数据标准粗略定义其元素特性。如美国国会图书馆数字报纸计划采用 METS 文档中的元数据对象描述模式(Metadata Object Description Schema, MODS)[15];芬兰国家图书馆历史报纸数字化项目主要参考 DC 标准描述报纸标题、出版商、出版日期等元素[16];中国国家图书馆民国数字化报纸描述采用 MARC 格式著录,主要记载民国报纸文献内容特征、载体形态、记录来源等信息[17]。上述内容均以参考成熟的元数据标准为主,并在项目实践中粗略揭示报纸资源特征,缺少对报纸资源内容层面的深度挖掘和标引,且各个资源库元数据描述较为单一化,尚未全方位描述报纸资源的语义特征及关联关系。

在报纸资源元数据描述理论研究方面,主要探讨元数据描述分析、元数据辅助用户交互检索[18]、识别用户检索模式、本地化元数据标准[19]等。如 J. H. Rho 对《朝鲜殖民报》进行详尽的元数据元素设计与应用,深入报纸知识内容单元,从报纸文章和广告元数据分析报纸属性,设计元数据标准[20],试图实现元数据

标准本地化目标，推动报纸资源长期保存；P. Fafalios以1987-2007年《纽约时报》为数据源，采用档案描述元数据和语义信息构建资源描述框架（Resource Description Framework，RDF）图，试图解决报纸档案资源的语义信息检索问题[21]；T. Bogaard等通过日志分析法探讨了荷兰国家图书馆历史报纸元数据在用户搜索行为方面的效用，识别用户的搜索模式[22]。国内学者则主要探讨历史报纸数字化、保存策略及报纸数据库建设等，缺乏对报纸资源的深层次语义描述和组织，且文献成果鲜少。代表性的有丁小蕾等参考DC标准从版次级和篇目级粗略设计了报纸元数据[23]；王静等从正文、广告、图片三大类型资源的元数据著录规则入手，重点阐述了该库的资源揭示与知识组织，并分析了《时报》数据库的功能构建情况[24]。以上2篇文献简要揭示了报纸资源的形式特征，缺乏语义深度。

综上所述，国内学者对近代报纸进行研究的文献量较少，尤其是图情档学科在报纸知识组织方面的研究成果稀缺，且研究深度不足。在新文科建设和数字人文浪潮的冲击下，近代报纸资源知识组织研究理应受到学界重视和关注。图情档学科也应发挥学科优势，对近代报纸资源进行全方位语义描述和揭示，充分发挥报纸应有的文献价值和史料价值。因此，为了弥补当前国内在该领域的研究空白，本文深入考量报纸资源特征，从本体和元数据描述两方面构建一个全面的近代报纸资源细粒度语义描述模型，并以《盛京时报》为例，采用RDF/XML语言实现资源的互操作和实践应用，推动近代报纸资源的高效组织及利用，提升近代报纸资源的知识组织能力及服务水平。

3 语义描述方法

当前学术界对知识进行语义描述的方法主要有两种：元数据标准和领域本体模型。成熟的元数据标准和领域本体可以为近代报纸资源语义描述模型构建提供参考借鉴。笔者通过梳理与近代报纸资源性质相似的元数据标准及领域本体，从而提炼出合适的部分进行复用，以此构建近代报纸资源细粒度语义描述模型。

3.1 元数据标准

元数据被称为数据的“数据”，是对基础数据进行更高维度和层次的抽象，由元素、修饰词及属性组成。元数据可以对数字信息资源进行内容属性和特征的描述，形成规范化数据描述体系，以便对资源进行有效管理、组织和检索。笔者对适用于近代报纸资源描述的元数据标准MARC、DC、EAD、MODS、CADAL和《古籍元数据规范》进行了梳理（见表1）。尽管上述元数据标准构成元素和元素限定词有所差异，但大多从资源内容属性、外部结构等方面对资源进行描述，其中DC、MODS描述范围非常广，普遍适用于各种网络信息资源；EAD主要用于描述档案和手稿资源，包括文本文档、电子文档、可视资料和录音资料等[25]，其高层元素由EAD头标、档案描述以及前置事项组成，其中EAD头标和前置事项提供检索信息，档案描述提供档案主体信息[26-27]；MARC主要用于图书馆书目数据描述；CADAL根据DC标准制定了报纸元数据著录规范，复用15个DC元数据，增加2个自定义元数据，版本信息（edition）和MARC记录，描述粒度较为粗糙；CDWA主要用于艺术品、收藏品等资源描述，包含分类、名称、创建者、时间、地点、相关作品等540个元素，描述非常全面；国家文物局制定的《古籍元数据规范》参考CDWA标准并自定义部分元素，共包括23个元素，描述较为精细。

表1 常用的元数据标准

元数据标准	简称	开发机构	发布时间	应用对象
Machine-Readable Cataloging[28]（机读编目格式标准）	MARC	美国国会图书馆	1970年	图书馆书目数据
Dublin Code[29]（都柏林核心元数据）	DC	联机图书馆中心、美国超级计算应用中心	1995年	网络信息资源
Categories for the Description of Works of Art[30]（艺术作品描述目录）	CDWA	艺术信息任务组	20世纪90年代初	艺术品、收藏品等资源描述
Encoded Archival Description[31]（编码档案描述）	EAD	美国档案工作者协会	1993年	档案和手稿资源
Metadata Object Description Schema[32]（元数据对象描述架构）	MODS	美国国会图书馆网络中心和MARC标准办公室	2002年	网络信息资源
China-America Digital Academic Library[33]（大学数字图书馆国际合作计划）-报纸元数据著录规范	CADAL	中国工程院、CALISH等	2002年	报纸资源
古籍元数据规范	-	中华人民共和国文物局	2017年	古籍类资源

3.2 本体模型

本体最早源于哲学领域,是对客观世界中事物的抽象概括。知识工程领域从哲学领域借鉴本体概念,并赋予了新的含义,被视为概念及概念之间关系的规范化和明确化描述,用来描述概念、属性和关系。R. Studer 等[34]认为本体是共享概念模型明确的形式化的规范说明。元数据和本体都是对信息资源的结构化描述方法。元数据主要对信息资源物理特征形态进行解释,旨在实现资源有效管理和检索。本体则侧重对知识进行描述,且可以揭示内容信息,如人、事、地、时、物等实体及实体概念之间的隐含关系。元数据是以资源为中心的辐射结构,本体则是去中心化的立体网状结构,元数据元素可作为本体中概念的属性[35]。常用的本体模型有 FRBR、BIBFRAME、CIDOC_CRM、FOAF 等。FRBR 是以"实体—关系"模型重构书目记录的功能需求框架。BIBFRAME 简化了 FRBR 模型,归纳出三组实体,目的是实现书目数据的关联发布。FOAF 是一种遵循 W3C 体系标准的资源描述框架词表,用于描述人与人之间的社会网络关系。CIDOC_CRM 采用面向对象方法定义了文化遗产领域实体(概念)、属性(关系),于 2014 年成为文化遗产领域国际标准(ISO21127:2014),其 2021 年 5 月版本包括 81 个类和 160 个属性。CRM 实体类型非常丰富,除了对文化遗产领域资源描述外,也适用于与文物相关的其他类型信息资源。因此,CRM 本体模型同样适用于近代报纸资源的描述建模。

4 近代报纸资源细粒度语义描述模型设计

4.1 近代报纸资源特征分析

近代报纸资源语义描述模型设计,需要全面考量近代报纸逻辑结构、物理布局和内容信息。本文以近代闻名中外的《盛京时报》为例,分析报纸资源相关特征,为近代报纸资源语义描述模型设计提供参考依据。《盛京时报》是日本人中岛真雄于 1906 年 10 月 18 日在沈阳创办的中文报纸,发行遍及东北、华北以南的一些城市甚至东南亚华语国家[36],于 1944 年停办。该报以国内时事和评论为主,主要汇聚了东北地区金融、商贸、交通、教育、文学等许多方面的信息,价值斐然,是研究东北军民抗日史、北洋军阀史以及中国近代史弥足珍贵的史料。

图 1 是 1906 年 10 月 25 日《盛京时报》内容,可以看出报纸的整体特征信息包括报名、版式、卷号、期号、版面、栏目等。图 1(a)、(d)版主要以"广告"为主,如"正金银行广告、三井洋行广告、延寿大药房广告"等,广告内容丰富,类型多样;图 1(b)、(c)版以"正文"为主,如社论、京师要闻、东三省要闻、各国要闻、专电、公文、市井杂俎及白话等栏目等均以文字形式描述内容,内容涉及"人、事、地、时、机构、职官"等实体,且类型多样化,社会新闻、时政新闻、文学小说等皆已呈现。

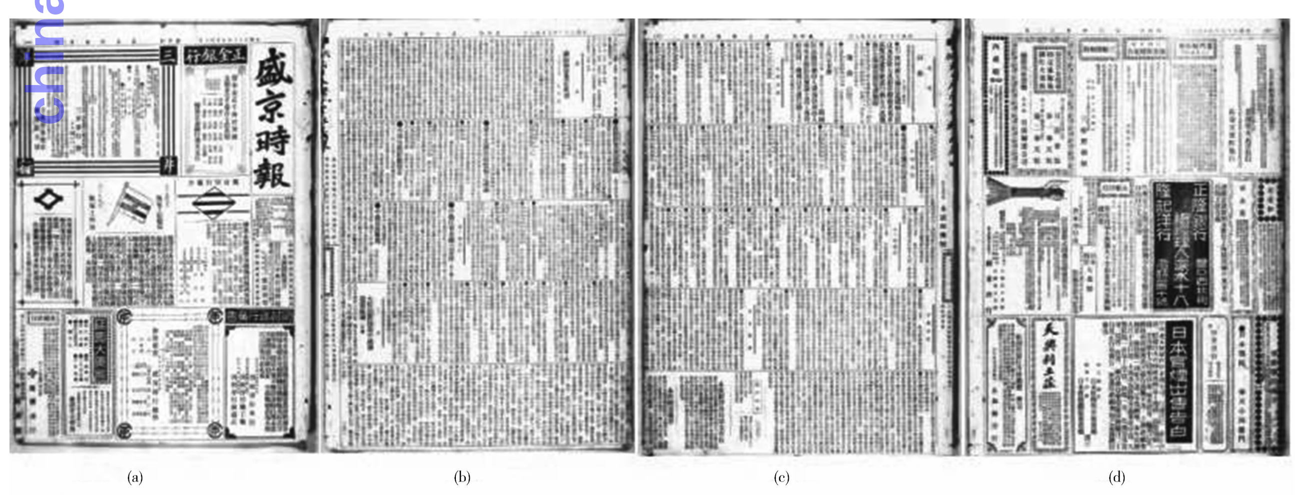


图 1 《盛京时报》(1906 年 10 月 25 日)

笔者为确保近代报纸资源语义描述模型的适用性,除了调研《盛京时报》外,浏览查阅了国家图书馆"中国历史文献总库·近代报纸数据库"中数百份报纸内容,总结归纳出报纸的整体形式特征(见表 2 元素列),并发现近代报纸在逻辑结构方面主要以"正文"和"广告"为主。因此,在设计近代报纸描述模型时,既要考虑到近代报纸的物理载体特征,又要兼顾报纸的逻辑结构信息,全方位解构近代报纸资源特征,抽取结构化信息。

表 2　近代报纸资源全局元数据描述

	元素	元素定义	元素复用标准	元素限定词	描述	示例
描述性元数据	title	题名	dc：title	alternative	报纸英文标题、日文标题、报眉标题，变更后报名等	盛京时报
	creation	创作	sach：creation	creator	创刊者	中岛真雄
				creationDate	创刊时间	1906.10.18
				creationPlace	创刊地	沈阳
	published	出版发行	sach：published	publisher	报纸出版者	盛京时报
				placeOfPublication	出版地点	沈阳
				issued	出版时间	1906
				printer	印刷者	－
				printedPlace	印刷地点	沈阳
				printedDate	印刷时间	1906
	contributor	其他贡献者	dc：contributor	－	社长、总主笔、总编辑、总经理、秘书、督印人等	菊池贞二
	type	报纸类别	dc：type	－	图片、文本、声音、视频等类型	文本
	language	内容语种	dc：language	－	报纸内容语种	Chinese
	description	描述	dc：description	abstract	以文本形式描述近代报纸相关信息，特别是其他元素未覆盖的信息，提要及各修饰词以外的附注说明可在此记录，如：对缺字的说明，对报名或者创作者的说明	该报收罗泛博，对当时我国内政、外交、经济、军事、文化、教育、社会风情等
	subject	主题	dc：subject	keywords	描述近代报纸的主题类型。编码体系修饰词可以采用中国分类主题词表	政治、军事、人文等
	column	栏目	自定义 np：colomn	－	报纸栏目信息	评论、东三省要闻、京师要闻、各省要闻、时论、批示、小说、社说、文苑、钦差行踪、专电、市井杂俎、公文汇录、广告
	publishedCycle	出版周期	自定义 np：publishedCycle	－	报纸出版周期有月报、周报、日报等	日报
	issue	总期数	自定义 np：issue	－	报纸总期数	12347 期
	materials	材质	sach：materials	－	近代报纸有油印、铅印、石印等材质	油印
	measurements	计量	sach：measurements	dimensions	用来描述近代报纸尺寸	27.0×20.0cm
				quantity	发行数量	－
管理性元数据	currentCondition	现状	sach：current Condition	levelOfCompleteness	描述近代报纸完残程度	缺/局部缺
				priority	描述保护优先等级，如修复级别	部分修复
	BHnum	编号	ead：num	－	报纸编号	－
	identifier	识别号	sach：identifier	generalRegistrationNumber	近代报纸在收藏管理时会有登记号、收藏号、排架号、分类号等作为标识	－
				otherLocalNumber		－
	currentLocation	所在位置	sach：current Location	geographicLocation	近代报纸入藏机构	辽宁省图书馆
				accessionDate	近代报纸入藏日期	－
	origination	来源	ead：origination	corpname	报纸来源机构	辽宁省图书馆
				persname	报纸来源机构	－
	relatedDigital Resources	数字对象	sach：related DigitalResources	digitalResourceIdentificationNumber	描述数字对象识别号	1906102800000001.jpg
				digitalResource RelationType	数字对象关系类型	原始图像
				digitalResource Format	文件格式	jpg
				digitalResource CreationDate	文件日期	－

（续表 2）

	元素	元素定义	元素复用标准	元素限定词	描述	示例
管理性元数据				digitalResource Creator	数字对象创建者	抗日战争与近代中日关系文献数据平台
				digitalResource Owner	数字对象所属机构	抗日战争与近代中日关系文献数据平台
				digitalResource Rights	数字对象权限	抗日战争与近代中日关系文献数据平台版权所有
				digitalResource Description	数字对象描述	盛京时报扫描图像
				digitalResourceLink	数字对象链接	https://www.modernhistory.org.cn/#/DocumentDetails_bz?fileCode=0008_bz_01000025
	rights	权限	dc：rights	accessRights	授权给谁访问	全部注册用户
				license	允许官方许可使用资源进行操作的法律文件	–
	relation	相关报纸	dc：relation	hasPart	包含在所描述资源的物理或逻辑中的相关资源	–
				hasFormat	相关资源，与已描述的资源基本相同，但采用另一种格式。	–

4.2 模型设计思路

近代报纸资源语义描述模型设计既要对报纸内容语义层面的相关实体进行识别和关联，如报纸所记载的人物、时间、地点、事件、机构、职官等相关实体信息，又要根据近代报纸逻辑结构对其物理层面的描述性信息进行细粒度挖掘，以此来充分描述近代报纸资源的整体特征。本文主要以《盛京时报》作为模型设计实例予以展示，以更清晰地展现模型所表达的知识要素及知识内在语义关系。

首先，确定模型的实体和关系。通过调研大量近代报纸资源析出相关实体类型，以实体为节点，谓词为连线，构建近代报纸资源实体之间的关联，并绘制模型示意图。

其次，定义模型的描述性信息，即描述性元数据或管理性元数据。既要对近代报纸资源进行全局描述，又要根据逻辑架构对内容信息进行局部描述。在析出近代报纸相关属性信息基础上，复用成熟的元数据标准，并自定义部分属性。

最后，形成一个完整的、高质量的、可互操作的、专指性强的近代报纸资源语义描述模型，实现近代报纸资源共享，推动近代报纸资源描述、定位、检索和组织能力。

4.3 模型设计过程

4.3.1 确定模型实体和关系

通过调研分析，发现《盛京时报》内容主要聚焦于人物、时间、地点、事件、机构、职官实体，且各类型实体之间关系紧密。本小节复用 CRM 本体中人物（Actor）、事件（Event）、地点（Place）、时间（Time-Span）实体，复用 ORG 本体中机构（Organization）实体，并自定义职官（Official）实体。具体内容如下：

人物（crm:Actor）。主要指《盛京时报》中所记载的人物，如孟宪彝、颜世清、柴田君、萩原君、宋春霆、德颐等。

事件（crm:Event）。主要指《盛京时报》中所记载的事件，如“日本在长春设立奉天总领事馆长春分馆”“长春开商埠”“日本建立满铁附属地”等具有重大影响意义的事件。

地点（crm:Place）。主要指《盛京时报》中所记载的空间位置或行政区划信息。如宽城子、长春、哈尔滨、沈阳、西三道街、头道沟、大连等。

时间（crm:Time-Span）。主要指《盛京时报》中所记载的时间点或时间段信息，一般以民国以前年号纪年方式居多，民国之后以公元纪年方式为主，此外还有农历纪年，如“光绪三十三年二月二十五”“本月初一日”等。

机构（org:Organization）。主要指《盛京时报》中所记载的机构信息，如“长春府”“长春领事馆”“巡警局”“民政司”“禁烟局”等实体机构。

职官（off：Official）。主要指《盛京时报》中所记载的职官信息，如“知府”“观察使”“太守”“总办”“领事”“局长”等。

实体关系是在实体类型基础上分析确定，以建立清晰准确的实体关系。本文以《盛京时报》中所记载的事件“督定永定河”为例，构建实体与关系示意图（见图 2）。CRM 中 E5 表示事件实体，该事件参与人物为 E21“孟宪彝”，事件发生时间为 E52“1916 年”，事

件发生地点为E53“永清县”，参与机构“顺直助赈局”，参与职官“永定河督办”；此外，对于人物实体E21“孟宪彝”，有子类实体E67“孟宪彝生年”和E69“孟宪彝卒年”，与E67和E69关联的实体有E52时间及E53地点，且人物所任职官为“知府”，所任机构为“长春府”，当然，人物职官及任职机构在不同的历史时期有所变化，此处仅展示其重要官职及任职机构信息。由此，《盛京时报》中有关人物、事件、地点、时间、机构、职官等实体及实体间关系构成了一个网络关联结构，对其语义关系予以揭示，实现报纸资源的知识组织与关联分析。

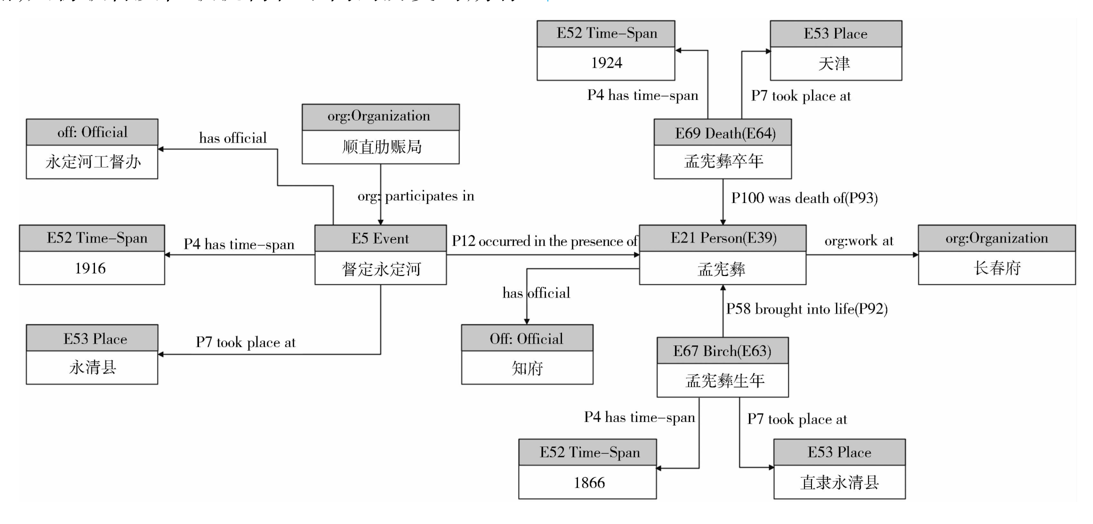


图2 《盛京时报》实体与关系示意

4.3.2 定义模型描述性信息

《盛京时报》实体与关系是针对报纸内容语义层面来构建语义描述模型。本节将以《盛京时报》为例，从物理层面特征进行描述性信息定义，以此构建一个完整的近代报纸资源细粒度语义描述模型。通过阅览《盛京时报》，笔者归纳总结其整体特征信息(见表2)。由于此部分特征提取工作均属于描述性元数据范畴，因此考虑复用成熟的元数据标准，并自定义部分元素，提取其特征信息。

(1)复用元数据标准说明。复用元数据需要充分考虑近代报纸资源特征，选择恰当适用的元数据标准。近代报纸区别于当代报纸，尤其在物理形态和刊载内容方面存在较大差异。近代报纸物理载体是以油印、铅印、石印等材质为主，而当代报纸大多以数字化文本呈现，且在出版周期、发行数量等方面也有别于当代报纸，因此在元数据描述时要考虑此要素；近代报纸刊载内容以中国近代半殖民地半封建的历史及中华民族奋起反抗的英勇事迹为核心，反映了中国近代社会的历史变迁历程，具有重要的档案价值和文物价值，在元数据描述时需考虑其档案及文物资源属性。

综上，本文主要从通用资源元数据标准、档案资源元数据标准及文物资源元数据标准3个维度探讨近代报纸资源描述规范。DC作为国际通用的元数据标准，具有简易性、灵活性和兼容性等特征，适用于广泛的网络信息资源描述，同样也适用于近代报纸资源描述，但元素专指性和针对性差，可作为近代报纸元数据描述补充框架；EAD作为档案资源元数据规范，主要用于描述档案和手稿资源。近代报纸也被视为历史档案的一部分，因此考虑复用EAD部分元素辅以模型描述，如档案物理形态方面元素(档案来源、档案编号)；《古籍描述规范》是国家科技支撑计划项目“文物数字化保护标准体系及关键标准研究与示范”课题研究成果，共发布62项标准规范，如文物、甲骨、舆图、壁画、拓片、古籍等元数据著录规范，适用于古籍、文物类资源描述。近代报纸具备文物资源属性，与《古籍描述规范》中对文物创作、文物出版发行、文物材质尺寸、文物数字对象和馆藏单位等元素描述相一致，因此考虑复用《古籍描述规范》部分核心元素。

(2)定义模型全局描述性信息。描述模型共包括21个元素，复用《古籍元数据规范》(用sach表示国家文物局(State Administration of Cultural Heritage))8个元素，复用DC标准(用dc表示)8个元素，复用EAD标准(用ead表示)2个元素，自定义(newspaper，用np表示)3个元素。具体信息见表2。

以《盛京时报》为例，绘制模型描述性信息示意图见图3，以进一步诠释说明表2信息。

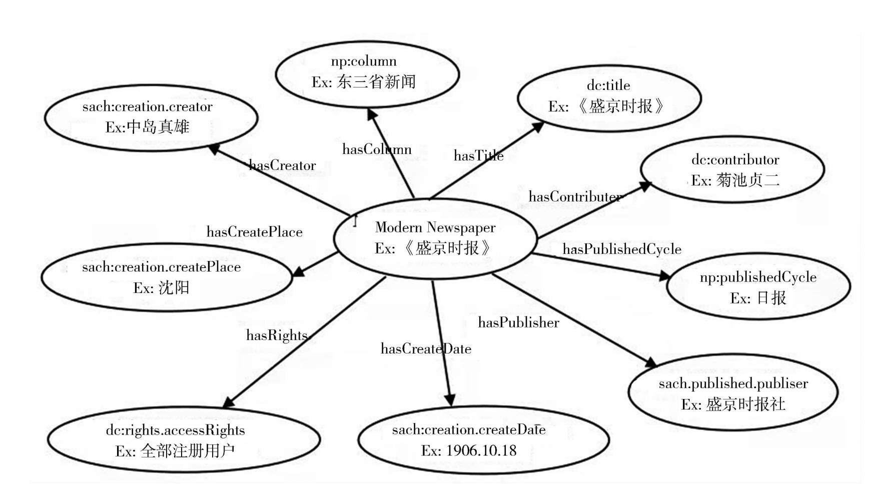


图 3 《盛京时报》全局描述性信息示意

(3)定义模型局部内容描述性信息。依据 3.1 内容,近代报纸资源主要以“正文”和“广告”两大逻辑体系为主。因此在描述近代报纸资源内容信息时,需要从这两方面着手,复用相关的元数据标准,并自定义部分元素,提取其特征信息。具体如表 3 所示,正文包括 3 个元素,复用 sach 标准 1 个、DC 标准 1 个、自定义 1 个。

以《盛京时报》为例,深入正文和广告内容层面,依据表 3 内容绘制示意图见图 4,以更清晰地呈现相关的特征信息。

表 3 近代报纸局部内容元数据描述

	元素	元素定义	元素复用标准	元素限定词	描述	示例
正文 Article	title	题名	dc: title	-	题名	“开放北满商埠电文”
	type	类型	dc: type	-	社会新闻、时政新闻、社论、文学作品	时政新闻
	event	事件	自定义 np:event	-	通知事件、抗议事件、谈判事件、战争事件、任命事件、罢免事件、抢劫事件……	通知事件
广告 Advertisement	title	题名	dc: title	-	题名	“阿稗精药片增食欲助消化治便秘”
	type	类型	dc: type	-	交通广告、金融广告、烟草广告、医药广告、银行广告、社会广告……	医药广告
	date	日期	dc: date	-	广告售卖日期	1937 年
	commodity	商品	自定义 np:commodity	name	商品名称	“阿稗精”
				company	商品生产公司	武田长兵衛商店股份有限公司
				category	金融、保险、洋行、铁路、药品、保健品、书籍等	药品
				price	商品价格	大瓶三个月分;中瓶四十五日分;小瓶半月分
				agent	商品代理商	奉天各大药房
				address	商品售卖地址	奉天春日町十三番地

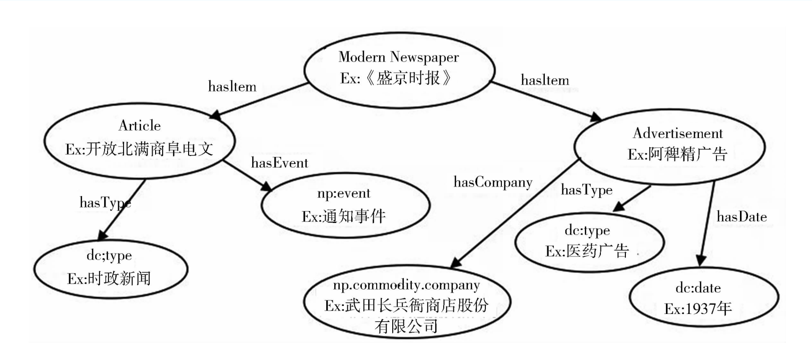


图 4 《盛京时报》正文和广告信息示意

综上分析，近代报纸资源语义描述模型从知识元细粒度视角出发，从内部语义关联到外部逻辑关联，解构近代报纸资源特征。首先，抽取人物、时间、地点、事件、机构及职官6类实体，并构建实体之间的关系，复用CIDOC-CRM模型构建逻辑关系示意图；其次，抽取模型全局描述性信息和局部内容描述性信息，并从近代报纸"正文"和"广告"两个维度剖析其内容特征，复用《古籍元数据规范》、EAD和DC标准构建语义关联示意图；最终形成一个清晰完整的近代报纸资源细粒度语义描述模型。

5 近代报纸资源细粒度语义描述模型应用

5.1 基于XML模型应用

为了更好地将近代报纸资源细粒度语义描述模型应用到具体的资源管理和存储中，实现不同元数据之间的互操作、数据交换和资源利用共享目标。本文采用XML(Extensible Marked Language)语言对其进行描述。XML是W3C推出的一种可扩展编辑语言，具有语言简洁、可扩展性高、互操作性强、方便网络传输等特性[37]，且为用户提供了灵活的标记扩展机制。本文将上述语义描述模型放置到XML中，采用Oxygen XML编辑工具创建xsd格式文档。Oxygen XML Editor是一款集XML查看和编辑功能为一体的软件，为用户提供XML创作和开发工具，可以自动完成代码校验、标签检测、代码高亮显示等功能。Oxygen XML涵盖所有XML标准，可扩展性高，支持连接大部分数据库。

本文使用Oxygen XML软件对描述模型进行编辑，构建Element和ComplexType，引入DC、EAD、sach标准，自定义标准np等，并创建命名空间(namespace)。Oxygen XML Editor具体操作界面和构建元素见图5。

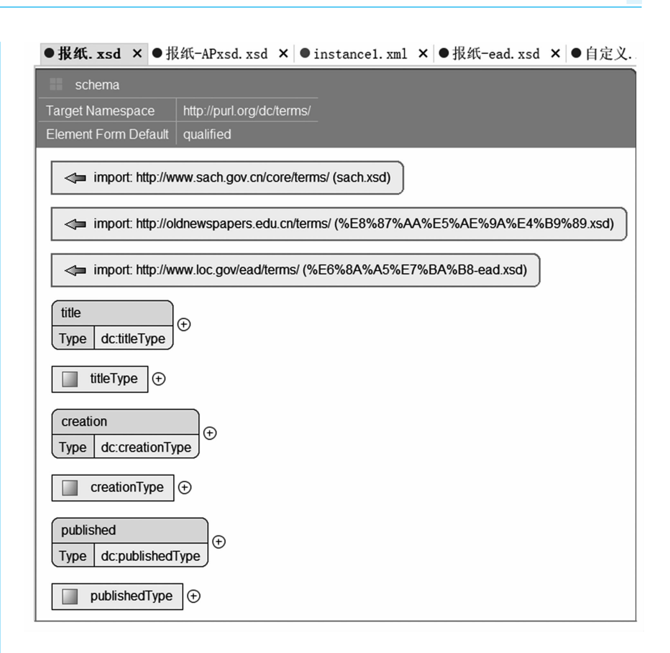


图5 Oxygen XML Editor操作界面

5.2 基于RDF/XML模型应用

在4.1基础上，进一步采用资源描述框架封装描述模型。RDF是W3C在XML基础上推荐的一种描述网络资源的标准[38]，用来对结构化元数据进行编码、数据交换和重用，为元数据提供一个可操作的载体和容器。RDF采用XML作为处理元数据的通用语法结构体系，为XML加入结构化约束提供清晰明确的语义表达方法[39]。RDF将资源看作对象，用统一资源定位符URI作为标识系统，并且提供一种RDF/XML的可扩展置标语言来书写和交换RDF模型。一般用三元组(主体(subject)、谓语(predicate)、对象(object))来表示资源对象。将XML与RDF结合，能充分发挥各自优势，实现资源的语义描述和应用。RDF可以引用不同的元数据方案，将多种元数据封装在统一的描述模型中，实现元数据之间的互操作。RDF规范性语法如下：每个RDF声明用一个rdf:Description元素表示，用rdf:about属性值声明主体的URI引用。词汇集使用本文所构建的近代报纸元数据描述框架，以《盛京时报》为实例予以应用，使同一资源的不同属性采用不同的元数据标准，实现不同元数据之间的互操作，更深层次地描述资源内容。

本文采用W3C RDF验证器[40]来验证RDF文档，结果见表4。从表4可以看出，W3C自动生成了《盛京时报》RDF数据模型三元组，即主体、谓语、对象。如主体为《盛京时报》，谓语为资源描述属性，对象则为属性值。RDF将不同的元数据进行封装，用三元组 <资源，属性，属性值> 灵活地描述报纸资源，且所有资源均通过唯一的URI来标识，使得资源以结构化方式呈现。本小节仅呈现《盛京时报》资源属性三元组，未来会进一步丰富本体描述模型，采用深度学习算法识别《盛京时报》中人物、事件、地点、机构、职官等实体，将实体存入到关系数据库中，通过外键设定实体关系，使用D2RQ工具将RDB关系数据转换为RDF数据格式，在此基础上借助Virtuoso数据库进行存储，并采用SPARQL语言进行检索，实现报纸资源的互联共享及利用。

表 4　RDF 验证结果

Number	Subject	Predicate	Object
1	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/Identifier	"http://oldnewspapers. edu. cn/shengjingshibao. html"
2	http://oldnewspapers. edu. cn/shengjingshibao. html	http://purl. org/dc/elements/1. 1/Title	"盛京时报"
3	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/creator	"中岛真雄"
4	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/creationDate	"1906 – 10 – 18"
5	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/creationPlace	"沈阳"
6	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/publisher	"盛京时报社"
7	http://oldnewspapers. edu. cn/shengjingshibao. html	http://oldnewspapers. edu. cn/terms/publishedCycle	"日报"
8	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/issued	"1906 – 10 – 18"
9	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/printer	"未知"
10	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/printedPlace	"沈阳"
11	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/printedDate	"1906 – 10 – 18"
12	http://oldnewspapers. edu. cn/shengjingshibao. html	http://purl. org/dc/elements/1. 1/contributor	"菊池贞二"
13	http://oldnewspapers. edu. cn/shengjingshibao. html	http://purl. org/dc/elements/1. 1/type	"文本"
14	http://oldnewspapers. edu. cn/shengjingshibao. html	http://purl. org/dc/elements/1. 1/language	"中文"
15	http://oldnewspapers. edu. cn/shengjingshibao. html	http://purl. org/dc/elements/1. 1/abstract	"《盛京时报》收罗泛博，对当时我国内政、外交、经济、军事、文化、教育、社会风情等，特别是对当时中国发生的重大事件，均有详略不等的报道；是研究近现代史、国际关系史、东北军民抗日史、北洋军阀史极为珍贵的资料，可供多方面的研究和利用。"
16	http://oldnewspapers. edu. cn/shengjingshibao. html	http://purl. org/dc/elements/1. 1/keywords	"东北近代史、近代报纸、北洋关系"
17	genid:A1311	http://www. w3. org/1999/02/22-rdf-syntax-ns#type	http://www. w3. org/1999/02/22-rdf-syntax-ns#Bag
18	http://oldnewspapers. edu. cn/shengjingshibao. html	http://oldnewspapers. edu. cn/terms/Column	genid:A1311
19	genid:A1311	http://www. w3. org/1999/02/22-rdf-syntax-ns#_1	"评论"
20	genid:A1311	http://www. w3. org/1999/02/22-rdf-syntax-ns#_2	"广告"
21	genid:A1311	http://www. w3. org/1999/02/22-rdf-syntax-ns#_3	"民国要闻"
22	genid:A1311	http://www. w3. org/1999/02/22-rdf-syntax-ns#_4	"东三省新闻"
23	http://oldnewspapers. edu. cn/shengjingshibao. html	http://oldnewspapers. edu. cn/terms/publishedCycle	"日报"
24	http://oldnewspapers. edu. cn/shengjingshibao. html	http://oldnewspapers. edu. cn/terms/issue	"12347 期"
25	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/materials	"油印"
26	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/dimensions	"27. 0 × 20. 0cm"
27	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/levelOfCompleteness	"缺/局部缺"
28	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/priority	"部分修复"
29	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. loc. gov/ead/terms/num	"未知"
30	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/geographicLocation	"辽宁省图书馆"
31	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. loc. gov/ead/terms/corpname	"辽宁省图书馆"
32	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. loc. gov/ead/terms/persname	"未知"
33	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/digitalResourceCreator	"抗日战争与近代中日关系文献数据平台"
34	http://oldnewspapers. edu. cn/shengjingshibao. html	http://www. sach. gov. cn/core/terms/digitalResourceDescription	"盛京时报扫描图像"
35	http://oldnewspapers. edu. cn/shengjingshibao. html	http://purl. org/dc/elements/1. 1/rights	"所有注册用户"

6　结语

近代报纸资源蕴藏的史料信息丰富，亟待知识组织技术和方法助力其开发。本文通过分析近代报纸资源物理布局、逻辑结构及内容信息，以《盛京时报》为例，从本体描述和元数据描述两个维度构建近代报纸资源细粒度语义描述模型，并采用 RDF/XML 语言应用描述模型。在本体语义描述中，采用 CRM 等本体概念模型表达《盛京时报》实体和关系，并以"督定永定河"事件为例，构建人物、时间、地点、事件、机构及职官实体的语义关系。在元数据描述中，结合近代报纸资源特征，复用 DC、EAD 和《古籍元数据规范》，构建报纸资源全局和局部描述性信息。

本文所构建的近代报纸资源细粒度语义描述模型具有理论和实践价值。在理论层面，拓宽了本体和元数据适用场域，将知识组织理论应用于近代报纸研究对象上，透过元数据揭示近代报纸资源物理载体特征，定位报纸资源位置信息，实现资源快速导航、发现以及多维语义检索；透过本体解析隐藏在报纸资源中的人物、时间、地点、事件、机构、职官等实体信息，并构建实体之间的语义关联关系，从细粒度视角挖掘知识单元，构造互联互通的语义网络。实践层面上，本文构建的语义描述模型同样适用于近代其他报纸资源的描述。同时也为当前近代报纸数据库在资源描述、检索及语

义服务等方面存在的不足提供参考路径，以推动近代报纸资源规范化管理和精细化服务，提升近代报纸资源利用效率，充分发挥近代报纸资源的史料价值和文献价值，进而传承社会记忆，发展中华民族优秀传统文化。未来会进一步采用深度学习模型对报纸不同类型实体进行识别，以此丰富实例内容。

参考文献：

[1] 张珂. 报纸起源和记者诞生及其演进发展的历史轨迹[J]. 陕西档案,2018(3):24-26.
[2] 杨雯,彭俊玲. 从出版文化遗产保护的角度看中国近代报刊的积累与开发[J]. 图书馆杂志,2015,34(10):78-84.
[3] 孔正毅. 试析中国近代报刊的历史文献价值[J]. 安徽理工大学学报(社会科学版),2009, 11(4):105-108.
[4] Voices and viewpoints in chronicling America: uses of historical news for education and outreach[EB/OL]. [2021-10-01]. http://library. ifla. org/id/eprint/1271/1/080-penn-en. pdf.
[5] NEUDECKER C, ANTONACOPOULOS A. Making Europe's historical newspapers searchable[C]//2016 12th IAPR workshop on document analysis systems (DAS). Santorini: IEEE, 2016: 405-410.
[6] Improving the discovery of European historic newspapers [EB/OL]. [2021-10-01]. http://library. ifla. org/id/eprint/1038/1/170-atanassova-en. pdf.
[7] 丁小蕾. 民国地方文献报纸数字化探索与实践——以首都图书馆地方文献民国报刊数字化为例[J]. 河南图书馆学刊,2016,36(12):98-100.
[8] 段晓林. 民国文献数据库开发现状研究[J]. 图书馆学研究,2016(20):42-45.
[9] KRAHMER A. Digital newspaper preservation through collaboration[J]. Digital library perspectives,2016,32(2):73-87.
[10] GEORGIEVA M. Successful management of an outsourced large-scale digitization newspaper project: tips for effective collaboration, increased productivity, and outstanding deliverables[J]. Journal of archival organization,2019,16(1): 52-74.
[11] 金彩虹. 民国时期四川地区报纸缩微胶片数字化——以四川大学图书馆为例[J]. 四川图书馆学报,2016(3):17-19.
[12] 陈桂香. 浅议民国报纸的数字化建设——以重庆图书馆为例[J]. 科技情报开发与经济,2013,23(4):27-29.
[13] JARLBRINK J, SNICKARS P. Cultural heritage as digital noise: nineteenth century newspapers in the digital archive[J]. Journal of documentation,2017,73(6):1228-1243.
[14] 肖红,槐燕. 民国报纸数字化实践中的质检问题探析[J]. 图书馆学研究,2017(7):61-78,87.
[15] MURRAY R L. Toward a metadata standard for digitized historical newspapers[C]//Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries. Denver: ACM Press,2005:330-331.
[16] AHONEN E, HYVONEN E. Publishing historical texts on the semantic web-a case study[C]//2009 IEEE international conference on semantic computing. Berkeley: IEEE,2009:167-173.
[17] 张玮. 民国报纸数字化验收常见问题研究——以国家图书馆为例[J]. 图书情报研究,2019,12(3):72-79.
[18] ALLEN R B, SCHALOW J. Metadata and data structures for the historical newspaper digital library[C]//Proceedings of the eighth international conference on information and knowledge management. Kansas: Association for Computing Machinery, 1999: 147-153.
[19] SEIFI L, AHMADZADEH N, PORDEL F. Digital preservation of old Persian periodicals in Iran with special reference to Iranian newspapers: strategies and challenges[C]//2015 4th international symposium on emerging trends and technologies in libraries and information services. Noida: IEEE,2015:81-85.
[20] RHO J H. Metadata elements design and application for Japanese newspaper 'Chosunsibo' issued in colonial Korea[J]. Journal of Korean Library and Information Science Society,2019,50(4):137-158.
[21] FAFALIOS P, KASTURIA V, NEJDL W. Ranking archived documents for structured queries on semantic layers[C]//Proceedings of the 18th ACM/IEEE on joint conference on digital libraries. Fort Worth: Association for Computing Machinery,2018:155-164.
[22] BOGAARD T, HOLLINK L, WIELEMAKER J, et al. Metadata categorization for identifying search patterns in a digital library[J]. Journal of documentation,2018,75(2):270-286.
[23] 丁小蕾. 民国地方文献报纸数字化探索与实践——以首都图书馆地方文献民国报刊数字化为例[J]. 河南图书馆学刊,2016,36(12):98-100.
[24] 王静,沈立力. 中文老报纸数据库的建设研究——以《时报》数据库建设为例[J]. 河南图书馆学刊,2018,38(10):106-108.
[25] 冯项云,肖珑,廖三三,等. 国外常用元数据标准比较研究[J]. 大学图书馆学报,2001(4):15-21,91.
[26] 赵屹. 网络档案信息检索的元数据设计[J]. 山西档案,2020(1):54-61.
[27] 宋欣,鲁国轩. 贝叶档案数字化建设中的元数据研究[J]. 浙江档案,2021(3):27-30.
[28] MARC[EB/OL]. [2021-10-01]. http://www. loc. gov/marc/.
[29] DC[EB/OL]. [2021-10-01]. http://dublincore. org/documents/dcmi-terms/.
[30] CDWA[EB/OL]. [2021-10-01]. http://www. getty. edu/research/publications/electronic_publications/cdwa/.
[31] EAD[EB/OL]. [2021-10-01]. http://www. loc. gov/ead/.
[32] MODS[EB/OL]. [2021-10-01]. https://www. loc. gov/standards/mods/.
[33] CADAL[EB/OL]. [2021-10-32]. http://cadal. edu. cn/index/home#page1.
[34] STUDER R, BENJAMINS V R, FENSEL D. Knowledge engineering: principles and methods[J]. Data and knowledge engineering,1998,25(1/2): 161-197.

[35] 夏翠娟. 文化记忆资源的知识融通:从异构资源元数据应用纲要到一体化本体设计[J]. 图书情报知识,2021(1):53-65.
[36] 郭广堃. 利用 TPI 系统建设特色馆藏数据库——以《盛京时报》数据库为例[J]. 河南图书馆学刊,2009,29(6):74-75.
[37] 任瑞娟,濮德敏,苗军民,等. 基于 XML/RDF 的 DC 元数据描述技术[J]. 情报杂志,2002(9):25-26.
[38] RDF[EB/OL]. [2021-10-01]. https://www.w3.org/RDF/.
[39] 鲁奎. 基于 XML/RDF 数字图书馆信息资源描述与应用研究[D]. 合肥:合肥工业大学,2003.
[40] RDF 验证器[EB/OL]. [2021-10-20]. https://www.w3.org/RDF/Validator/.

作者贡献说明:
孙绍丹:论文撰写、修改;
邓君:提出研究思路,论文撰写、修订;
常严予:资料查找;
张子姝:协助论文资料查找;
沈涌:论文校对。

Design and Application of the Fine-Grained Semantic Description Model of Modern Newspaper Resources: Taking *Shengjing Times* as an Example

Sun Shaodan[1] Deng Jun[1] Chang Yanyu[1] Zhang Zishu[1] Shen Yong[2]
[1] School of Business and Management, Jilin University, Changchun 130012
[2] School of Public Health, Jilin University, Changchun 130022

Abstract: [**Purpose/Significance**] This paper designs a scientific and standardized fine-grained semantic description model of modern newspaper resources, and reveals the characteristics and relationships of modern newspaper resources in depth, in order to provide references for the effective management, organization, knowledge discovery and knowledge service of modern newspaper resources. [**Method/Process**] By analyzing the logical structure, physical layout, contents of modern newspaper resources, starting from two aspects of domain ontology description and metadata description, this paper reused the CIDOC-CRM ontology conceptual model, EAD, DC and *Ancient Book Metadata Specification*, and took *Shengjing Times* as an example to design a fine-grained semantic description model for modern newspaper resources. Then, it used Oxygen XML tool to describe the semantic model in RDF/XML language to realize element interoperability and model application. [**Result/Conclusion**] This paper aims to provide an operable fine-grained semantic description model for the modern newspaper resource organization, provide a basic guarantee for the construction of the modern newspaper resource database, the standardized newspaper management and the application system development, and promote the development, utilization and sharing of the modern newspaper resources.

Keywords: modern digital newspaper resources semantic description model *Shengjing Times* RDF/XML